Data based correction of MC

Antoni Aduszkiewicz

University of Houston

PPFX meeting Feb 19, 2021

Introduction

- Investigate two methods of correcting simulation based on data:
- With data interpolation:
  1. Interpolate data in fine (p, θ) bins, preserving integrals in the original data bins from the publication
  2. Divide interpolated data by MC in fine bins $\rightarrow$ correction factor
  3. Draw spectrum in $(x_{\mathrm{F}}, p_{\mathrm{T}})$ bins, weighting each track by the correction factor
- Without data interpolation:
  1. Divide data by MC in original bins $\rightarrow$ correction factor
  2. Draw spectrum in $(x_{\mathrm{F}}, p_{\mathrm{T}})$ bins, weighting each track by the correction factor
- Test data: π^- spectra produced in p+p interactions at 80 GeV/c generated with EPOS and VENUS models, 5M interactions each.
  In this presentation I pretend EPOS is "data" and VENUS is "MC"

Procedure without data interpolation

- Correction factor is obtained by dividing data by MC in the same (p, θ) bins as in which the data was published:

$$c(p, \theta) = \frac{\mathtt{data}(p, \theta)}{\mathtt{MC}(p, \theta)} \tag{1}$$

- Data spectrum in $(x_\mathrm{F},\ p_\mathrm{T})$ can be obtained by filling histogram with the same MC tracks using corrections obtained in the previous step:

$$\mathtt{data}(x_\mathrm{F}, p_\mathrm{T}) = \mathtt{MC}(x_\mathrm{F}, p_\mathrm{T}) \cdot c(p, \theta) = \mathtt{MC}(x_\mathrm{F}, p_\mathrm{T}) \cdot \frac{\mathtt{data}(p, \theta)}{\mathtt{MC}(p, \theta)} \ . \tag{2}$$

- The procedure takes *shape* of the MC distribution and scales it by a data-based correction factor in each bin.

Procedure with data interpolation

- Interpolating data $\mathtt{data}(p, \theta)$ results in data in fine binning $\mathtt{data}^f(p, \theta)$
- Correction factor is obtained by dividing it by MC in fine binning:

$$c(p, \theta) = \frac{\mathtt{data}^f(p, \theta)}{\mathtt{MC}^f(p, \theta)} \tag{3}$$

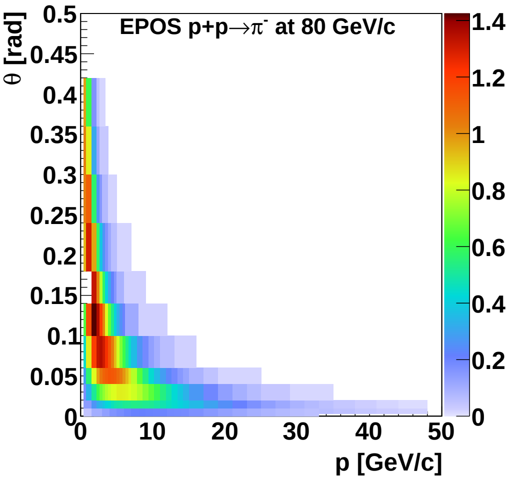
- Data spectrum in $(x_\mathrm{F},\ p_\mathrm{T})$ can be obtained by filling histogram with the same MC tracks using corrections obtained in the previous step:

$$\mathtt{data}(x_\mathrm{F}, p_\mathrm{T}) = \mathtt{MC}(x_\mathrm{F}, p_\mathrm{T}) \cdot c(p, \theta) = \mathtt{MC}(x_\mathrm{F}, p_\mathrm{T}) \cdot \frac{\mathtt{data}^f(p, \theta)}{\mathtt{MC}^f(p, \theta)} \ . \tag{4}$$

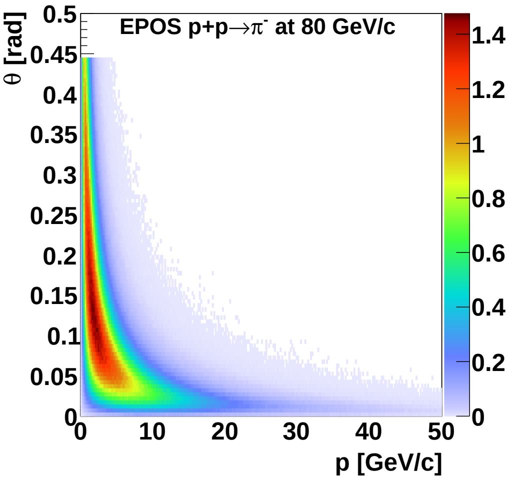
Q: **Why do we need MC at all?** Note that even very fine bins in (p, θ) space can grow large in $(x_\mathrm{F}, p_\mathrm{T})$ space. If we didn't use MC it would be equivalent to assumption that the data is perfectly flat within the fine bin

EPOS “data”

data ($d^2n/\mathrm{d}p\mathrm{d}\theta$ [$(\mathrm{GeV}/c)^{-1}$]) in NA61 paper bins

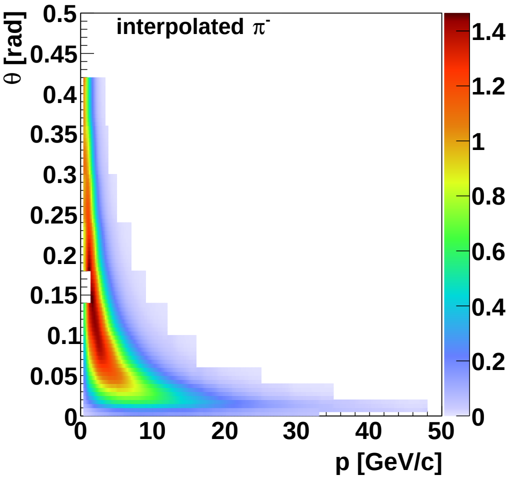


data in fine bins (in real world we can’t see it)

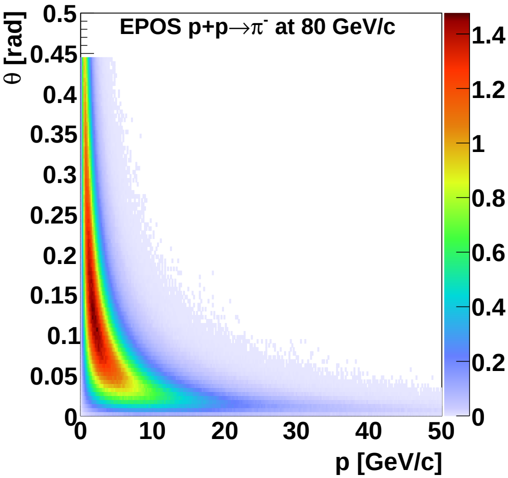


EPOS “data” interpolated

interpolated data

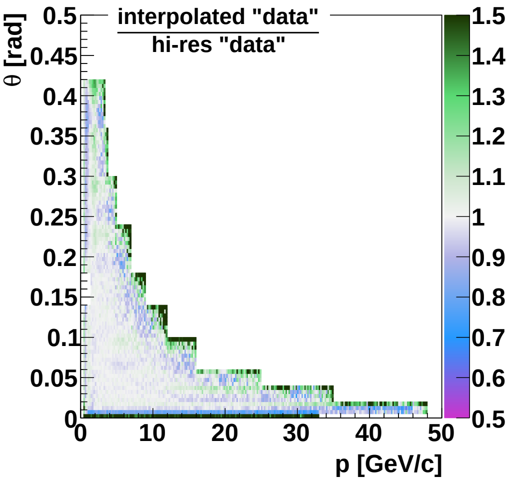


data in fine bins (normally we can’t see it)

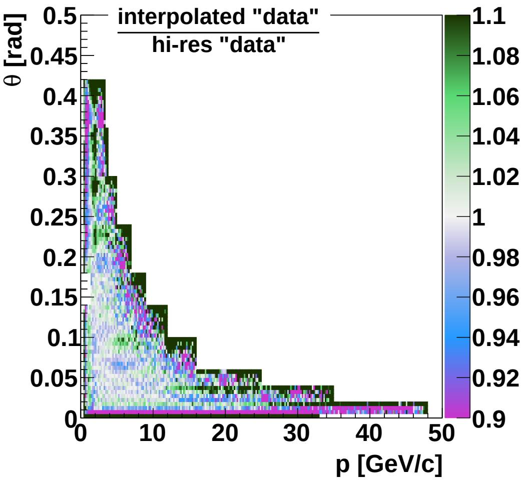


Interpolation error

(interpolated data) / (data in fine bins)

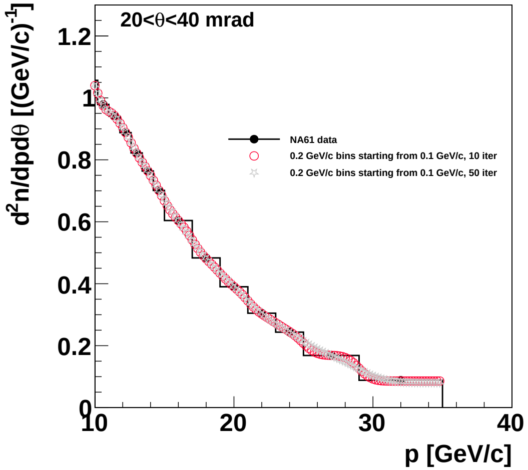


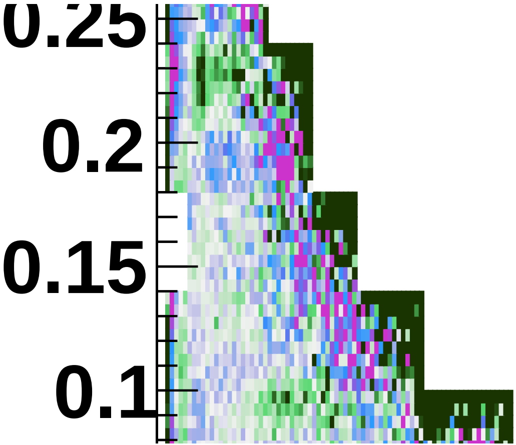
same plot, color scale zoomed



- Large errors at the edges, 10% biases in the whole range
- Of course the errors would be much larger with no interpolation. Also the integral is preserved

What happens at the edges?

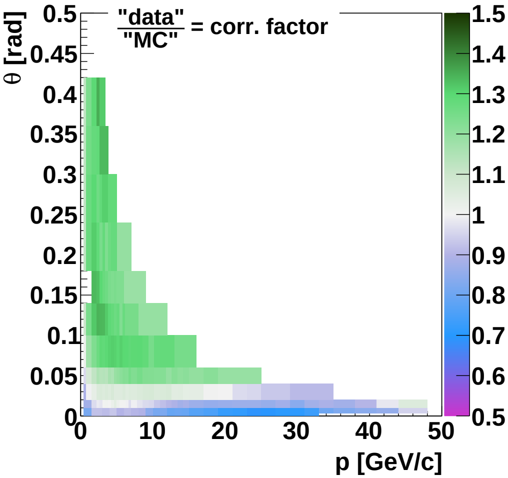




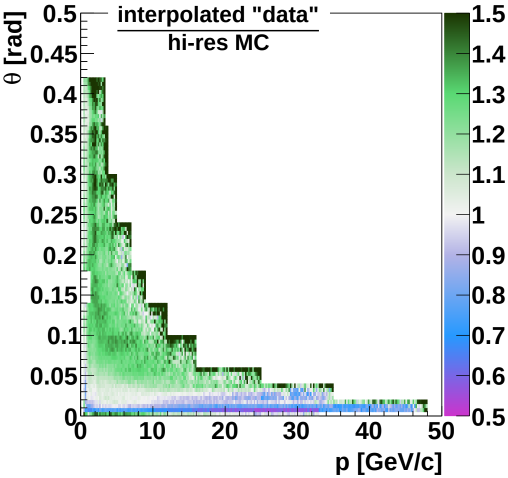
- There is no data to interpolate at the edges, so we begin to extrapolate. Extrapolation requires assumptions.
- Notice how much less error is in the “dent” area at $\theta = 0.15$ rad

Correction factor

correction in original bins (no interpolation)

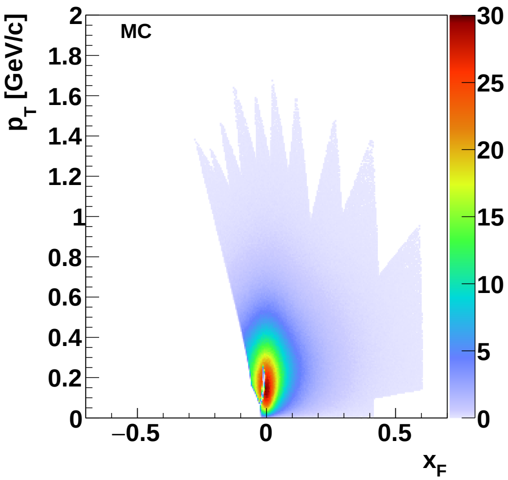


correction based on interpolated data

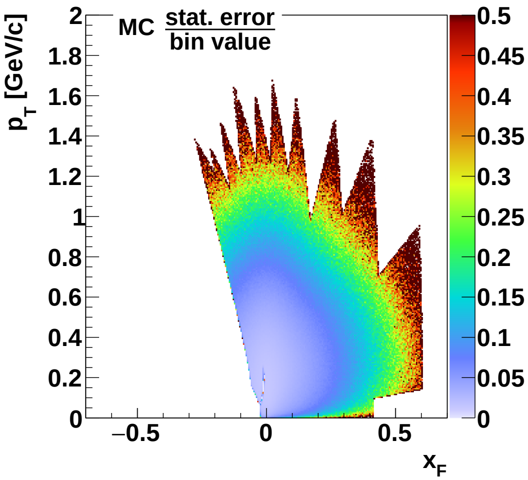


MC in (x_F, p_T)

MC spectrum

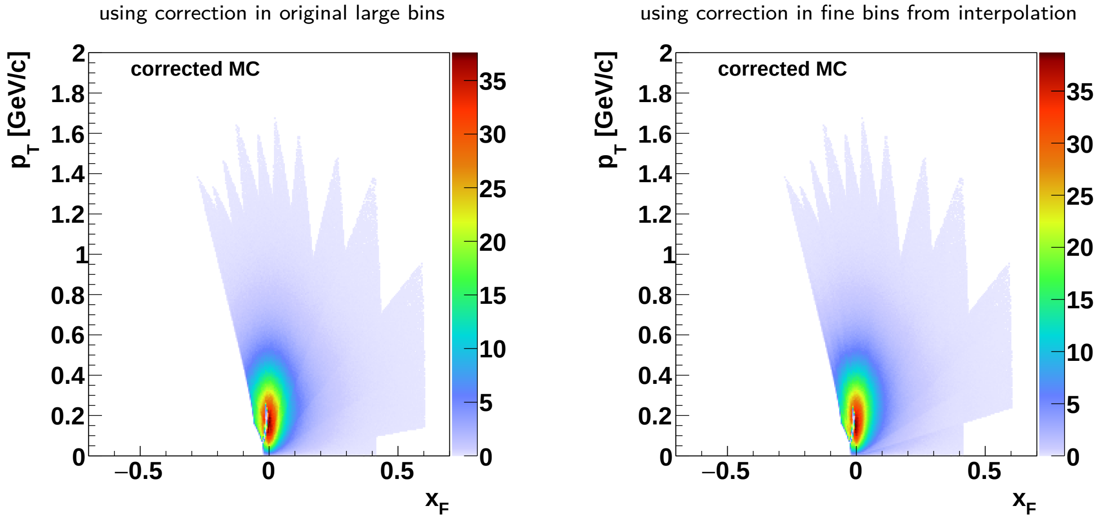


relative statistical uncertainty of MC



- Plotted only in the region covered by the "data" bins
- Statistical errors significant at the edges

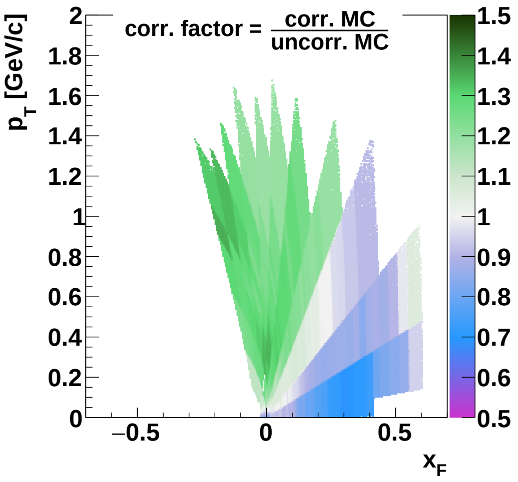
Corrected MC in (x_F, p_T)



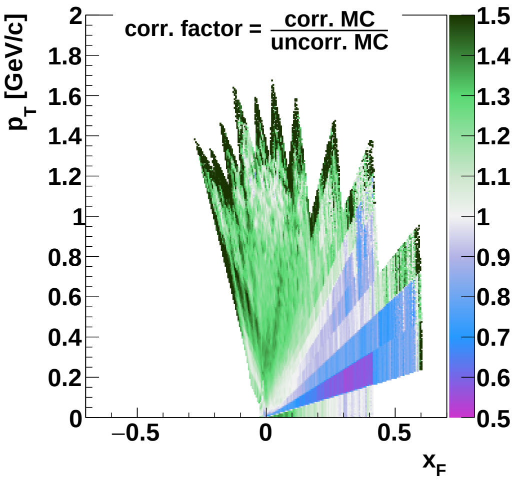
- For each particle a correction weight was applied based on its (p, θ) coordinates

Effective correction factor

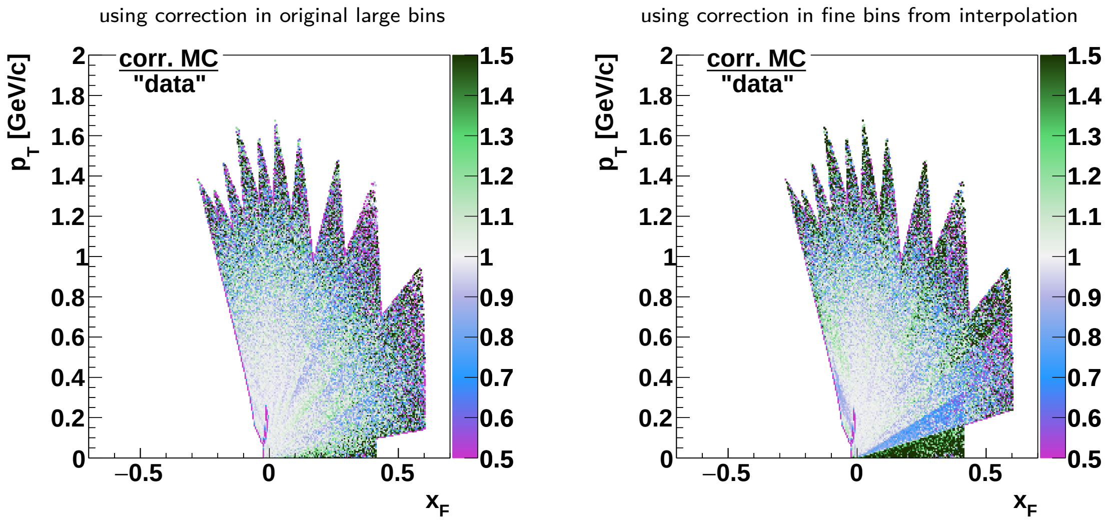
using correction in original large bins



using correction in fine bins from interpolation



Bias of corrected data



- Statistical fluctuations need to be disregarded in these plots

Summary

Comparison of two methods

- Both method introduce some 10% systematic biases here and there $\rightarrow$ expected as we have data in large bins only
- Interpolation introduces large errors close to some edges

Possible improvements

- Interpolation method
  - Improve the interpolation method $\rightarrow$ tried already, little improvements
  - Omit bins at the edges from the analysis $\rightarrow$ waste of data
  - Manually add fake data bins at the edges to improve interpolation $\rightarrow$ lot's of work, introduces model dependence, difficult to defend
- No interpolation method
  - Interpolate the correction factor $\rightarrow$ 1. I'm not yet sure how 2. risk of running into the same issues with interpolation 3. But possibly interpolating the correction factor introduces less error than interpolating the spectrum?
- Ask model creators to tune their models
  - They may have much better experience in solving these kind of problems than us

Other future steps

- Each method requires more testing with various test data sets to estimate size of bias it introduces